Machine Learning 1

Lecture 5.4 - Supervised Learning
Classification - DecisionTheory

Erik Bekkers

(Bishop 1.5)

Slide credits: Rianne van den Berg

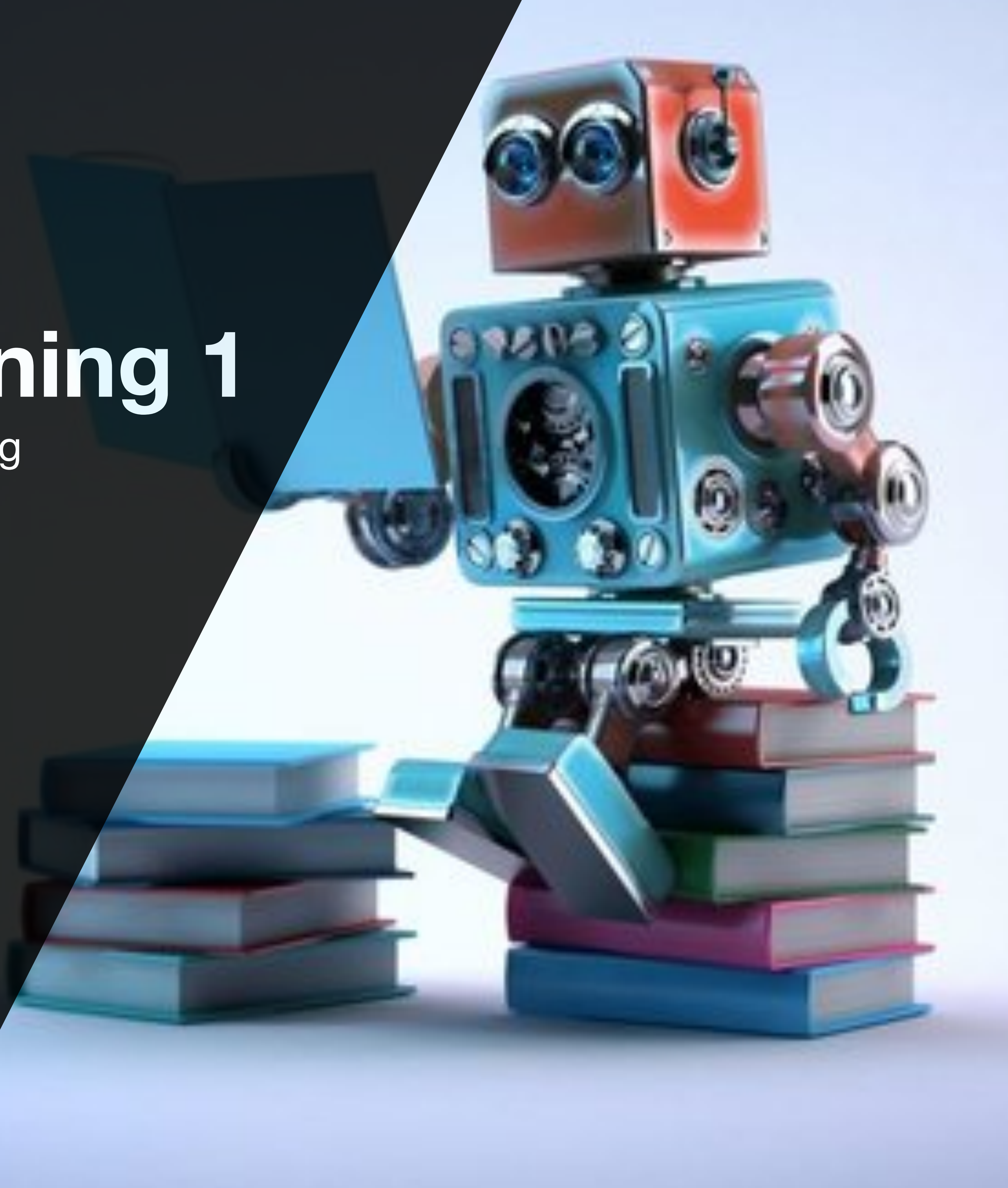


Image credit: Kirillm | Getty Images

Decision theory

- Dataset: Input vectors $\underline{x} \in \mathbb{R}^D$, ground truth targets $t \in \{C_1, \dots, C_K\}$
- Divide input space $\mathbb{R}^D$ into K decision regions $\mathcal{R}_k$, $k = 1, \dots, K$
- Every observed datapoint $\begin{cases} \text{ground truth} & t_n = C_j \\ \text{prediction} & \hat{t}_n = C_k \quad (\underline{x} \in \mathcal{R}_k) \end{cases}$
- Confusion matrix: ground truth classes vs. predicted classes

$$\begin{array}{c|c} & \begin{array}{cccc} \mathcal{R}_1 & \mathcal{R}_2 & \dots & \mathcal{R}_K \end{array} \\ \begin{array}{c} C_1 \\ C_2 \\ \vdots \\ C_K \end{array} & \begin{pmatrix} 6 & 1 & \dots & 0 \\ 5 & 3 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 2 & 0 & \dots & 8 \end{pmatrix} \end{array}$$

$\leftarrow$ given by classifier $\hat{t}_n = C_k$

ground truth $\rightarrow$

- Diagonal elements: correctly classified
- Off-diagonal elements: misclassified

Decision theory: Misclassification Rate

- Classification goal: Minimize the misclassification rate
- Assume observations are drawn from joint distribution $p(\underline{x}, t)$
- Probability of a misclassification:

$$p(\text{mistake}) = \sum_{i=1}^{K} \sum_{k \neq i} p(\mathbf{x} \in R_i, C_k)$$

$$= 1 - \sum_{k=1}^{K} p(x \in R_k, C_k)$$

Minimizing misclassification rate

- Assign x to class C_k if $p(\underline{x}, t = C_k) > p(\underline{x}, t = C_j), \; j \neq k$
- Note: $p(x, C_k) = \underline{p(C_k|x)}p(x)$ Check for the largest post. class prob

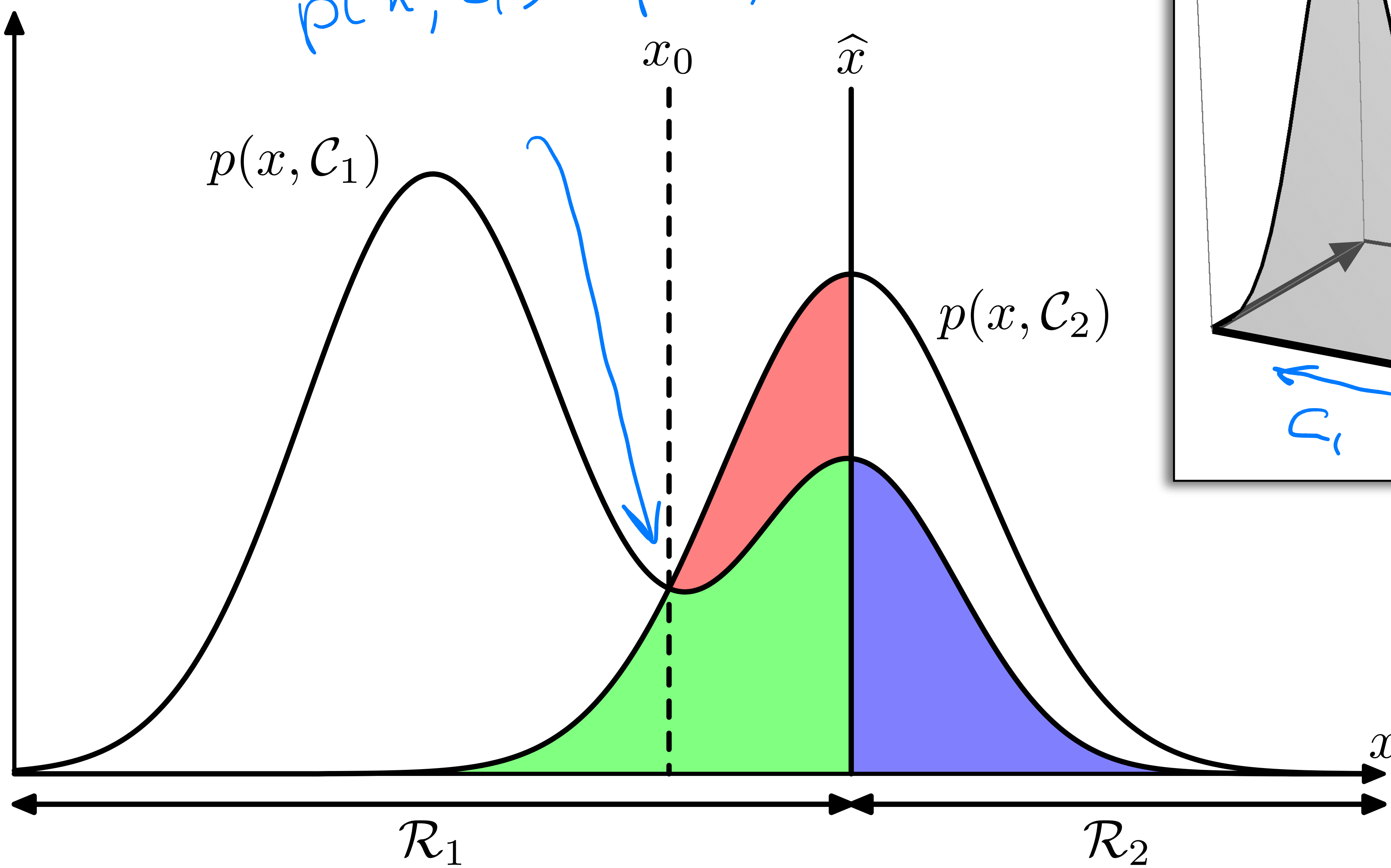
$p(C_k | x) > p(C_j | x), \; j \neq k$

Decision theory: Misclassification Rate

$\hat{x}$: Decision boundary

x_0 : Optimal Decision boundary

$p(x, C_1) = p(x, C_2)$



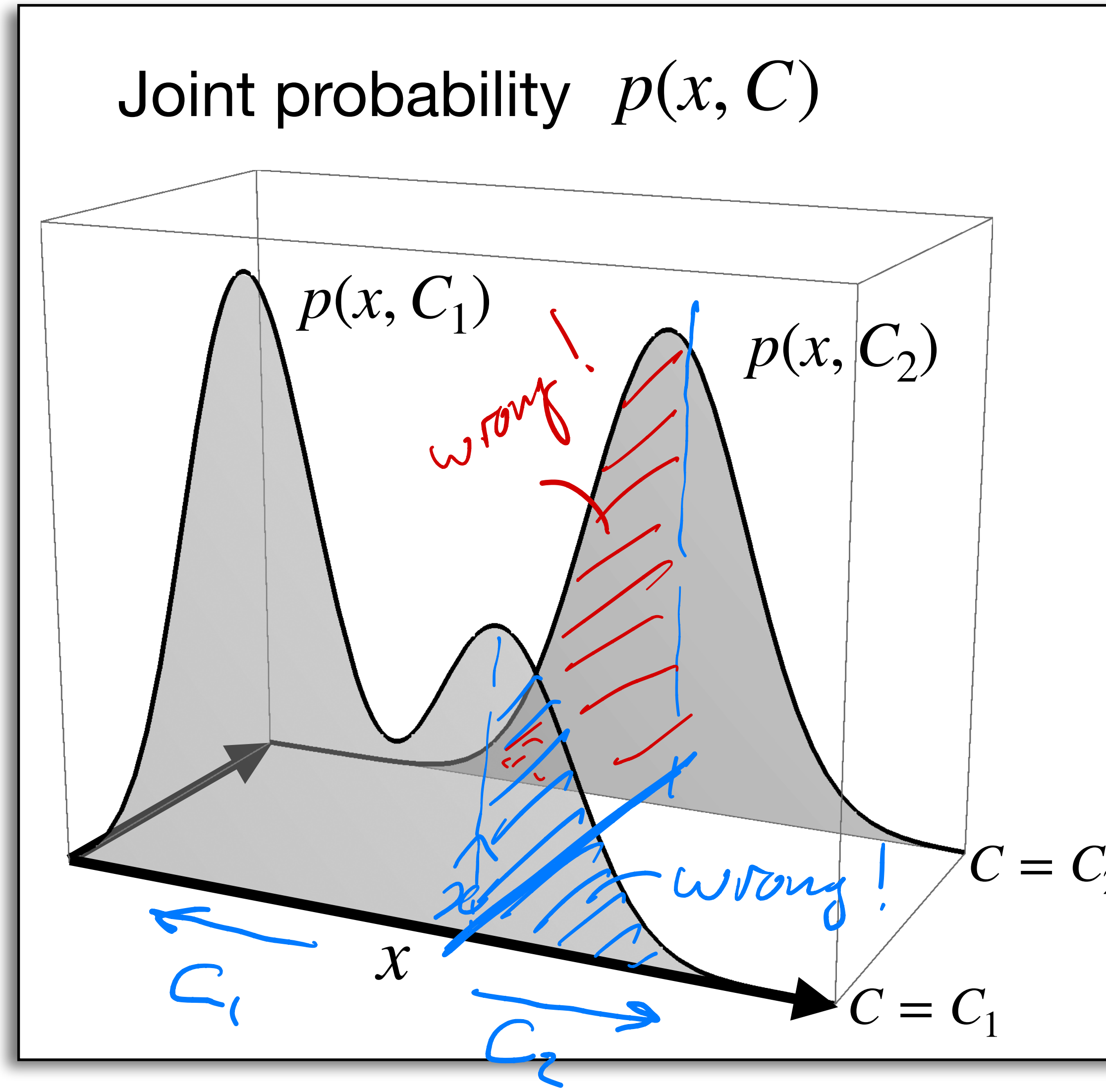


Figure: joint probability distributions and decision boundary (Bishop 1.24)

Minimizing the Misclassification Rate: Problems

1
- Not all errors have the same impact!

<u>Example: Medical diagnosis of cancer</u>

- Error 1: Label a healthy person as having cancer.
- Error 2: Label a sick person as healthy. Lack of treatment!

- If cancer only occurs in 1% of all patients, a classifier which labels everyone as healthy has a misclassification rate of 1%!

2 Class imbalance

Expected Loss

- Possible solution: use different weights for different error types

$$
\begin{array}{ccc}
 & \text{label cancer} & \text{label healthy} & \\
L = & \begin{pmatrix} 0 & 1000 \\ 1 & 0 \end{pmatrix} & & \begin{array}{c} \text{true cancer} \\ \text{true healthy} \end{array}
\end{array}
$$

- Expected loss: $\mathbb{E}[L] = \sum_{k,j} L_{kj} \int_{\mathcal{R}_j} p(x, C_k) \mathrm{d}x$

error $j \neq k$

Minimize expected loss:

- Assign x to C_k if $\sum_{j=1}^{K} L_{jk} p(x, C_j)$ is minimal

Classification Strategies

1
- *Discriminant functions*

Direct mapping of input to target $t = y(\underline{x}, \underline{w})$

2
- *Probabilistic discriminative models*

Posterior class probabilities: $p(C_k | x)$

3
- *Probabilistic generative models*

Class-conditional densities: $p(x | C_k)$

Prior class probabilities: $p(C_k)$

1. $p(x, C_k) = p(x|C_k)\, p(C_k)$
2. $p(C_k|x) = \dfrac{p(x|C_k)\, p(C_k)}{p(x)}$